DOCUMENT RESUME

ED 052 234 TM 000 638

AUTHOR Stake, Robert E.

TITLE Measuring What Learners Learn (With a Special Look at Performance Contracting).

INSTITUTION Illinois Univ., Urbana. Center for Instructional Research and Curriculum Evaluation.

SPONS AGENCY Illinois State Office of the Superintendent of Public Instruction, Springfield.
PUB DATE 71
NOTE 41p.

EDRS PRICE EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS Criterion Referenced Tests, Educational Accountability, *Educational Objectives, Educational Testing, Evaluation Criteria, Evaluation Techniques, Grade Equivalent Scores, *Learning, *Measurement, *Performance Contracts, Performance Criteria, Performance Specifications, Student Testing, Teacher Evaluation, *Testing Problems

ABSTRACT

A discussion of performance contracting, defined as an agreement between a group offering instruction and a school needing the services, is presented. Four major hazards to direct measurement of specific learning are considered: poor statement of objectives; selection of the wrong tests; misinterpretation of test scores; and depersonalization of contemporary life. These and other problems such as human and testing error, valid criterion testing, and the question of when to test, are discussed in full. The relationship of these hazards of performance measurement to performance contracting, and to regular school programs, is presented. (AG)

MEASURING WHAT LEARNERS LEARN*
(With a Special Look at Performance Contracting)

Robert E. Stake
Center for Instructional Research and Curriculum Evaluation
University of Illinois at Urbana-Champaign

"Can there be teaching if there is no learning?" Hear again one of the lines from the educator's catechism. The question is not to be taken literally. Good teaching, elegant teaching, without student benefit, of course is possible--though doubly wasteful. The question is rhetorical. Professionals and laymen alike sanctify that teaching-learning contract that results in better student performance.

Measuring the learning is no small problem. Teachers, as a matter of course, usually are able to observe that individual students are or are not learning. Sometimes they cannot. And increasingly, outsiders are reluctant to take the teacher's word for it. Gathering "hard-data" evidence of student learning is a new and ominous challenge. Of course, we have tests. But the results of our testing have seldom been adequate grounds for the continuing faith we have in education.

Present Demands. Expectations of testing are on the rise because schools have been told to be accountable--to demonstrate publicly what they are accomplishing (Lieberman, 1970; Bhaerman, 1970). Increasing educational costs and increasing frustration with social and and political problems have brought higher demands for answers to an important question: What are we getting for our education dollar?

Educators have been challenged to become more explicit and more functional in lesson plans and school budgets; to identify the gains and losses

*A paper prepared with financial support from the National Educational Finance Project and the Office of the Superintendent of Public Instruction, State of Illinois.

children make in reading, singing, and the many human talents; and to realize that the events of the classroom are not unrelated to the events of the street, the marketplace, and City Hall (Cohen, 1970). Educators have been told to learn about systems analysis, operations research, cost-benefit analysis, program planning and budgeting, and other models for orderly and dispassionate treatment of institutional affairs (Lessinger, 1970).

Some critics of contemporary education are bothered greatly by the fact that educational practice is so intuitive, impulsive, inefficient, and resistant to change. Others continue to be bothered more by passionate but naive efforts to substitute technical procedures for personal attention. Thorndike (1921), Tyler (1950), and Krathwohl (1969) have been persuasive advocates of a more rational, explicit, performance-oriented school. But Atkin (1968), Oettinger (1969), and Dyer (1970) have cautioned that formal analyses and production models can be narrow, irrelevant, and even oppressive. It is safe to say that all specialists in testing and instruction believe that it is possible to measure many specific educational outcomes and to use such measurements in improving educational decisions. But a few of these same specialists are among the most vehement critics of present testing (Glaser, 1963; Grobman, 1971).

Tests for Performance Contracts. The performance contract is an agreement between a group offering instruction and a school needing services (Lennon, 1971). Reimbursement is to be made in some proportion to measured student achievement. Especially for children having special needs, such as nonreading, handicapped, or gifted children, a new way of getting special instruction is appealing. A "hard-data" basis for evaluating the quality of instruction is appealing. In performance contracting student gains are the criterion of successful teaching.

In the first federally sponsored example of performance contracting for the public schools, Dorsett Educational Systems of Norman, Oklahoma, contracted to teach reading, mathematics, and study skills to over 200 poor-performance juniors and senior high school students in Texarkana, Texas. Commercially available, standardized tests were used to measure performance gains.

Are such tests suitable for measuring specific learnings? To the person not intimately acquainted with educational testing it appears that performance testing is what educational tests are for. The testing specialist knows that this is not so. These tests have been developed and administered to measure correlates of learning, not learning itself.

Most tests are indirect measures of educational gains, correlates of learning rather than direct evidence of achievement. Correlation with important general learning is often high, but correlation of test scores with performance on many specific educational tasks is seldom high. Tests can be built for specific competence, but there is relatively little demand for them and many of them do a poor job of predicting later performance of either a specific or general nature. General achievement tests "predict" better. The test developer's basis for improving tests has been to work toward better predictions of later performance rather than better measurement of present performance. Assessment of what a student is now capable of doing is not the purpose of most standardized tests. Especially when indirect-measurement tests are used for performance contracting, but even with direct-assessment tests, errors and hazards abound.

In this paper I will identify the major obstacles to direct measurement of the specific things that learners learn.

The Errors of Testing

Answering a National School Board Journal (November 1970) questionnaire on performance contracting, a New Jersey board member said, "Objectives must be stated in simple, understandable terms. No jargon will do and no subjective goals can be tolerated. Neither can the nonsense about there being some mystique that prohibits objective measurement of the educational endeavor." Would that our problems would wither before stern resolve. But neither wishing nor blustering rids educational testing of its errors. They exist.

Just as the population census and the bathroom scales have their errors, educational tests have theirs. The technology and theory of testing are highly sophisticated; the sources of error are well known (Lindquist, 1951; Cronbach, 1969). Looking into the psychometrist's meaning of "A Theory of Testing," one finds a consideration of ways to analyze and label the inaccuracies in test scores (Lord, 1952). There is mystique, but there is also simple fact: No one can eliminate test errors. Unfortunately, some errors in testing are large enough to cause wrong decisions about individual children or about school-district policy.

The whole idea of educational testing is thought to be an error by some educators and social critics (Hoffman, 1962; Holt, 1969; Silberman, 1970; Sizer, 1970). Bad social consequences of testing, such as the perpetuation of racial discrimination (Goslin, 1970) and pressures to cheat (McGhan, 1970) continue to be discussed. But, as would be expected, most test specialists believe that the promise in testing outweighs these perils. They refuse responsibility for gross misuse of their instruments and findings; and they concentrate their attention on reducing the errors in specific tests and test programs (Lennon, no date).

Some technical errors in test scores are small and tolerable. But some testing errors are intolerably large. Today's tests can, for example, measure vocabulary word-recognition skills sufficiently accurately. Today's tests cannot adequately measure listening comprehension or the ability to analyze the opposing sides to an argument.

Today's test technology is not refined enough to meet all the demands put on it. The tests are best when the performance is highly specific--when, for example, calling for the student to add two numbers, recognize a misspelled word, or identify the parts of a hydraulic lift. When a teacher wants to measure performances calling for the higher mental processes (Bloom et al, 1956), such as generating a writing principle or synthesizing a political argument, our tests give us scores that are less dependable. See Table 1 for several examples.

Table I

Examples of Items of High and Low Validity
in Conventional Standardized Achievement Tests

High validity--"basic mental process" items:

*1. Which one of the following phrases about wave motion defines period?

a. the maximum distance a particle is displaced from its point of rest
b. the length of time required for a particle to make a complete vibration
c. the number of complete vibrations per second
d. the time rate of change of distance in a given direction

*2. Directions: In each group below, select the numbered word or phrase which most nearly corresponds in meaning to the word at the head of that group, and put its number in the parenthesis at right.

( ) antelope

a. fruit b. animal c. prelude d. feeler e. gallop

*3. The first movement of a sonata is distinguished from the others by:

a. rapidity and gaiety
b. length and complexity
c. emotional abandon
d. sweetness and charm
e. structural formality

4. Which of these would help you decide whether or not you used the word "filter" correctly in a sentence?

a. encyclopedia
b. dictionary
c. thesaurus
d. English grammar textbook

Table I (continued)

Lower validity--"higher mental process" items:

*5. A and B were arguing about the desirability of adopting a nationwide system of compulsory health insurance in the United States. B said that, while he had no fundamental objection to health insurance, he felt strongly that people should not be compelled to participate in it. "Now look here," he said, "Do the people want health insurance or don't they? I don't think they do, but in either case, compulsory insurance is bad. If the people really want health insurance, there is no need for compulsion. If they don't want it, it is impossible to force them to participate. So the answer is clear."

Which of the following statements most nearly expresses the logical conclusion of B's argument?

a. Health insurance is bad.
b. Compulsory health insurance is bad.
c. Compulsion is impossible.
d. Compulsion is unnecessary.
e. Compulsion is either unnecessary or impossible.

**6. Directions: In each situation below, you are given introductory information about a person's action or conclusion. This is followed by several independent statements of evidence. Decide whether the added information in each statement makes it more or less probable that the action or conclusion is correct. For each statement, mark the answer space under a if the added information makes it more probable that the conclusion is correct; under b if the added information makes it less probable that the conclusion is correct; under c if the added information makes it neither more nor less probable that the conclusion is correct.

Situation: I predict that our team will win the basketball tournament next week. With the exception of one player, our team is the same as last year when we won easily. Furthermore, we have a 13-3 won-lost record this season.

Statements:

a. Another team in the tournament has been undefeated against substantially the same teams.
b. Our closest competitor will be relying mainly on sophomores to carry it to victory.
c. The first game will be played Monday morning instead of Monday afternoon as previously announced.

*From Bloom et al, 1956; reproduced here with permission.

**From Analysis of Learning Potential, Form A, 1970, published by Harcourt Brace Jovanovich, Inc.; reproduced here with permission.

Unreached Potentials. Many educators feel that the most human of human gifts--e.g., the emotions, the higher thought processes, interpersonal sensitivity, moral sense--are beyond the reach of psychometric testing. Most testing specialists disagree. While recognizing an ever-present error component, they believe that anything can be measured. The credo was sounded by E. L. Thorndike (1918):

> "Whatever exists at all exists in some amount. To know it thoroughly involves knowing its quantity as well as its quality. Education is concerned with changes in human beings; a change is a difference between two conditions; each of these conditions is known to us only by the products produced by it--things made, words spoken, acts performed, and the like. To measure any of these products means to define its amount in some way so that competent persons will know how large it is, better than they would without measurement. To measure a product well means so to define its amount that competent persons will know how large it is, with some precision, and that this knowledge may be conveniently recorded and used. This is the general Credo of those who, in the last decade, have been busy trying to extend and improve measurements of educational products.
>
> "We have faith that whatever people now measure crudely by mere descriptive words, helped out by the comparative and superlative forms, can be measured more precisely and conveniently if ingenuity and labor are set at the task. We have faith also that the objective products produced, rather than the inner condition of the person whence they spring, are the proper point of attack for the measurer, at least in our day and generation.
>
> "This is obviously the same general creed as that of the physicist or chemist or physiologist engaged in quantitative thinking--the same, indeed, as that of modern science in general. And, in general, the nature of educational measurements is the same as that of all scientific measurements."

Testing men believe it still. They are not so naive as to think that any human gift will manifest itself in a 45-minute paper-and-pencil test. They believe that, if given ample opportunity to activate and observe the examinee, any trait or talent or learning that manifests itself in behavior can be measured

with reasonable accuracy. The total "cost" of measuring may be a hundred times that of administering the usual tests, but they believe it can be done. The final observations may rely on professional judgment, but it would be a reliable and validated judgment. The question for most test specialists then is not "Can complex educational outcomes be measured?" but "Can complex educational outcomes be measured with the time and personnel and facilities available?"

If we really want to know whether or not a child is reading at age-level, we have a reading specialist listen to him read. She observes his reading habits. She might test him with word recognition, syntactic decoding, and paragraph-comprehension exercises. She would retest where evidence was inconclusive. She would talk to his teachers and his parents. She would arrive at a clinical description which might be reducible to such a statement as "Yes, Johnny is reading at or above age-level."

The scores we get from group reading tests can be considered estimates of such a clinical judgment. These test scores correlate positively with the more-valid clinical judgments. Though more objective, such estimates are not direct measurements of what teachers or laymen mean by "ability to read." . ... . Achievement gains for a sizable number of students will be poorly estimated by them. It is possible that the errors in group testing are so extensive that--when fully known--businessmen and educators will refuse to accept them as bases for contract reimbursement.

Professional Awareness. Classroom teachers and school principals have tolerated standardized test errors (as much as they have) because they have not been obligated to make important decisions on the basis of test scores alone. Actually, it is seldom in day-to-day practice that they use test scores (Hastings, Runkel, and Damrin, 1961); but, when they do, they use them, in combination with

other knowledge, to estimate a child's progress in school and to guide him into an appropriate learning experience. They do not use tests as a basis for assessing the quality of their own teaching.

In performance contracting the situation is supposed to be drastically changed. Tests are indicated as the sole basis for contract reimbursement. The parties must decide how much to pay the contractor for instructing each child. An error in testing means money misspent. Graduation and reimbursement decisions are to be made without reliance on the knowledge and judgment of a professional observer, without asking persons who are closest to the learning (i.e., the teacher, the contractor, the student) whether or not they see evidence of learning. They are to be made entirely by objective and independent testing. The resulting human errors and technical misrepresentations will be numerous. On the following pages I will discuss four major hazards: (1) attending to the wrong objectives, (2) selecting the wrong tests, (3) misinterpreting the test scores, and (4) adding to the depersonalization of contemporary life.

Choice of Objectives

I am addressing this paper to the measurement of objectives already specified. It is important to recognize that at no time--in any real educational practice--are instructional objectives completely and finally specified.

No statement of objectives is final. Changes in aim, as well as changes in priority, occur throughout training even in the more highly structured instructional programs. Some people feel that this is what is wrong with much classroom instruction: It cannot pick a target and stay fixed on it. But other people are convinced that classroom instruction is too fixated, too inflexible, that teachers are too unwilling to adapt to the changing goals of students and society.

No statement of objectives says exactly what it ought to. Every statement has its ambiguity; each word can be misunderstood; we cannot expect any list to say exactly what its authors want it to. Verbal statements of objectives cannot perfectly represent human purpose. All this does not mean that educators should not state their objectives, but it does mean that educators should continue to look for better ways of representing their objectives. They should expect them to change from beginning to end of semester and beyond. They should regard any statement as an approximation. Objectives remain in flux, never completely free of misrepresentation by our tests and observations, in even the most stable curricula.

Specification Benefits. Identifying the goals of education in formal, rational terms is recognized as a powerful way to change professional practice (Tyler, 1950; Mager, 1962). To recognize that objectives will change is not to argue that they should not be stated in advance of training. An awareness of

purpose by both teacher and student is usually desired. Only occasionally will an educational experience be highly successful if there is no advance expectation as to what should occur. Usually the activity will be improved if the opportunity to learn is deliberately provided for. Often instruction will be improved if lesson plans focus on desired behaviors rather than entertain spontaneous interests and distractions.

Outside evaluation of the success of instruction is made much simpler and possibly more effective by the prespecification of objectives. Popham (1969) has identified these and other benefits that accrue to those who state their instructional objectives in advance and stick to them.

Specification Costs. But each of these possible benefits carries with it a cost. Stating objectives properly is a lot of work. Some other possible costs are less obvious. In each of the next six paragraphs I will identify an important cost that may be incurred in specifying objectives prior to training.

To specify what is to be accomplished always fails to represent the sum total of what is desired. Language fails to portray exactly what we want. The error may be small and unimportant, or it may not. But to some extent there will be a misrepresentation of purpose.

The singularity of any list of objectives--even if it has 100 separate objectives--disregards the disparity in what teachers, students, and citizens need and want. In a pluralistic society, different people have different priorities. Gooler (1971), for example, found that teachers put more emphasis on humanistic curricular objectives than parents do. In his article in the Saturday Review last fall, Schrag (1970) said:

> "Any single, universal public institution--and especially one as sensitive as the public school--is the product of a social quotient verdict. It evaluates the lowest common denominator

> of desires, pressures, and demands into the highest public virtue. It cannot afford to offend any sizable community group be it the American Legion, the B'nai B'rith, or the NAACP."

Publicized statements of objectives are likely to represent nobody's objectives.

Any public display of educational goals evokes political and social reaction (Lortie, 1967). Educators--as other people--are seldom candid in the face of hostile criticism. They are likely then to state (and possibly emphasize in the classroom) objectives that are less controversial. Pressure to state objectives is transformable into pressure to change objectives.

The schools presently pursue many more objectives than any educator can specify, more than he chooses to admit (Gooler, 1971). The results of a specification of objectives, for good or ill, is to increase substantially the emphasis on some objectives and to decrease substantially the emphasis on others. Some objectives are more easily specified and more easily measured than others. It is almost certain that easy-to-measure objectives will get increased emphasis when a statement of objectives is drawn up.

The language of behavioral specification is such that behavioral processes (recalling, solving, writing, observing, etc.) are given greater emphasis as to what the school will do; and subject matter (the Civil War, use of quotation marks, conservation of energy, the nature of knowledge, etc.) will get less. Gagné has claimed (1967, p. 21) that subject matter is preserved to any desired extent by behavioral objectives; but the AAAS Elementary Science Curriculum--its creators relied heavily on his counsel--is a curriculum which attends relatively little to the traditional categories and relationships of science. Increased emphasis on performance is likely to bring decreased emphasis on content.

Furthermore, when curricular objectives are spelled out in advance, it is more difficult for a teacher to seize an opportunity to teach something the students are obviously ready for and wanting to learn (Atkin, 1968). And it is more difficult for a teacher to assign needed remedial work when the schedule, and perhaps the syllabus, call for "completion" of specific units.

The listing of trade-offs could go on. There are many things that happen when you try to state educational objectives in "simple, understandable terms." McNeil (1967), Jenkins and Deno (1969), Eva Baker (1970), and Zahorik (1970) carried out empirical studies to examine the good and bad effects of specification and planning. Improved student performance on the specified objectives in some circumstances appears to be attributable to the specification itself. But Zahorik found that planning resulted in less attention to immediate concerns of the pupils. More research on the overall effects of specification is needed. For each effort to identify more specifically what will be learned, to identify it earlier, and to identify it formally as a statement of instructional objectives, it seems that there are both potential benefits and hazards for the ongoing instructional process.

Criterion Testing Procedures

Among test developers the most vexing problem has always been "the criterion problem," the problem of correlating test scores to a true criterion. For validating a new test, the developer needs to ascertain that at least for a small, carefully measured reference group of students there is a high correlation between what the test measures and what is already known about that group that the test is supposed to indicate. A high correlation signifies that for that criterion chosen, the test is valid. The high scorers on a study-skills test, for example, would be the students who independently and by direct observation are judged to have the best study capabilities. True criterion observations--whatever the criterion might be--are not readily available on most students. Because of the difficulty and expense, any one standardized test will be validated against only one, or a very few, criterion variables. The most common criterion variable is a course grade given by a teacher or a grade-point average.

For performance testing, the standardized test--the right, already-validated, standardized test--is not likely to exist. The purposes of the contracted-for instruction are relatively sharp, e.g., to increase reading speed and comprehension--and the available tests have been validated against a more general criterion, e.g., grades in reading. The educator has a choice between using a not-quite appropriate available test and building an expensive and questionably valid test. The problem is a vexing one: how to select or construct the appropriate items, observations, or test to serve as the criterion of learning for the purpose of the contract.

Three questionable aspects of the criterion test need careful thought. There is (1) a question about relying on performance as a criterion indicator of

benefit from instruction, (2) a question about measuring complex performances with simple tests, and (3) a question of "teaching for the test." The first two are related to hazards in the choice of objectives as described in the previous sections.

Savings. An objection to the performance test is that it does not reveal one of the outstanding benefits of instruction: savings. In learning-research jargon, "savings" is the increased ease of relearning something just because it was studied before. Whether or not a student learns something to mastery level, he usually forgets some or all of it. When he needs to know it, in school or out, he usually has an opportunity to relearn it. Immediate recall is just not as important as test designers assume.

It is usually much easier to learn the second time than the first. It it, of course, easier to learn on that later occasion than it would have been had the learner not studied the lesson before. Sometimes it is easier because the learner knows how to go about learning it the second time. This savings is an important benefit from instruction. Learning how and when to use reference sources for particular topics is a major, but poorly recognized, instructional objective. Such learning shows up as savings. Savings and long-range retention are among several things,* in addition to immediate retention performance, that should be looked at in deciding whether or not instruction deserves reimbursement.

*Others: (1) improving typical as well as maximal behavior; (2) developing awareness of contexts where special skills are needed; (3) increasing structure and "organizers" for learning; (4) provision of opportunity to learn, (5) increasing desire to learn; (6) use of good adult models; and (7) treatment of students with dignity and humanity, etc. Perhaps the school officials should be paid a bonus if they identify an appropriately broad set of objectives or fined if they do not.

Complex Performances. It is unrealistic to expect that a project director can find or create paper-and-pencil test items, administrable to large numbers of students in an hour's interval by persons untrained in psychometric observation and standardized diagnostics, objectively scorable, valid for purposes of the performance contract, and readily interpretable. The more complex the training, the more unrealistic the expectation. One manner of compromise is to substitute criterion test items measuring simple behaviors for those measuring the complex behaviors targeted by the training. For example, the director may substitute vocabulary-recognition test items for reading-comprehension items or knowledge of components in place of actual dismantling of an engine.* The substitution may be sound, but the criterion test should be validated against performances directly indicated by the objectives. It almost never has been.

It would be unrealistic to expect that the benefits of instruction will be entirely apparent in the performances of learners at test-taking time. The tests to be used probably will evoke relatively simple behavior. Ebel (1971) said:

> "...most achievement tests...consist primarily of items testing specific elements of knowledge, facts, ideas, explanations, meanings, processes, procedures, relations, consequences, and so on."

He went on to point out that more than simple recall is involved in answering even the simplest vocabulary item.

Much more complex behavior is needed for answering a reading-comprehension item. An example of an excellent reading-comprehension item, from the Iowa Tests of Basic Skills, is shown in Table 2. The items here are clearly

*Tendencies to teach for the test in this situation must be checked.

Table II

A Passage and Ten Questions to Measure Reading Comprehension

PARAGRAPH 1 When your teacher says "O.K.," you know that all is well. Do you know how we happen to use two letters of our alphabet for words? Do you know the words for which the two letters stand?

PARAGRAPH 2 The custom of using O.K. to mean that all is correct is now 100 years old. It began during the election year of 1840. William Henry Harrison, a candidate for president, came to Urbana, Ohio, to make a speech. A large number of people went out to meet him. When they returned to town, one of the wagons carried a large banner on which was written, "The people is oll korrect." The spelling, of course, was wrong; the sign meant "all correct."

PARAGRAPH 3 The enemies of General Harrison made fun of the poor spelling of his friends. Harrison's friends, however, used the saying to advertise their candidate. They said he was the candidate of the common man. Since many people of that day could not spell well, saying that Mr. Harrison's friends could not spell made him still more popular with the common people. Soon, instead of saying "oll korrect," people were saying just "O.K.'

PARAGRAPH 4 After the election Daniel Leffler, an innkeeper of Springfield, Ohio, put a sign over the door of his house which read, "The O.K. Inn." This inn was on the great national road. Many people stopped to eat and many others saw the strange sign as they drove by. Harrison had been elected president, and people remembered the "Oll Korrect" and the "O.K." of the election. The sign on the inn kept the memory alive. Besides, the food at the inn was "oll korrect" as advertised. People began to say "O.K." when things were right.

QUESTIONS ON SELECTION NO. 1

0. What two letters are mentioned in the first line?
   1) A.B. 2) O.K. 3) X.Y. 4) P.S.

1. Why did Mr. Harrison come to Urbana?
   1) To see the "oll korrect" sign
   2) To get people to vote for him
   3) To visit some old friends
   4) To stay at the O.K. Inn

2. What is the purpose of paragraph 1?
   1) To tell where the expression "O.K." originated
   2) To get the reader interested in the article
   3) To tell that letters can be used for words
   4) To ask the reader if he knew what O.K. meant

3. Who first used the letters "O.K." for the words "Oll Korrect"?
   1) Wm. H. Harrison
   2) Daniel Leffler
   3) The innkeeper
   4) The article does not give his name

4. What is the topic of paragraph 2?
   1) The origin of the term "O.K."
   2) The election of Harrison
   3) The poor spelling of Ohio men
   4) The meaning of the term "O.K."

5. What is the author's purpose in paragraph 4?
   1) To show that the innkeeper's sign developed from the misspelled campaign banner
   2) To show that the innkeeper's sign helped make O.K. a popular expression
   3) To tell about the location and food of "The O.K. Inn"
   4) To tell who first used the term "O.K."

6. Why did the misspelled words on the banner make Harrison more popular?
   1) They made people believe Harrison was "oll korrect," as the banner said
   2) They suggested that Harrison was a common man
   3) The innkeeper of a popular inn put them on his sign
   4) Most people disliked good spellers

7. Why was Daniel Leffler mentioned?
   1) Because he helped Harrison win the election of 1840
   2) Because he carried a banner saying, "The people is oll korrect"
   3) Because he owned an inn that had very good food and service
   4) Because he helped to make O.K. a common expression

8. What is the author's purpose in writing this article?
   1) To make the reader curious by asking questions
   2) To show why it is sometimes good to misspell words
   3) To show how the English language has developed in the last 100 years
   4) To tell how one of our common expressions started

9. What kind of speller was Mr. Harrison?
   1) Good
   2) Average
   3) Poor
   4) The article does not give any clue

10. The expression "O.K." probably would have been forgotten if one of the following were true. Which one is it?
   1) If Daniel Leffler had been a better speller
   2) If the people had known who Daniel Leffler was
   3) If Mr. Harrison had been an unpopular candidate
   4) If Mr. Harrison's enemies had been good spellers

Printed by permission from Iowa Every-Pupil Tests of Basic Skills, Test A, Form N, Advanced Battery, 1942.

calling for more than the literal meanings of the words read. The student must paraphrase, interpret--what we expect readers to be able to do.

These items and ones for problem solving and the higher mental processes do measure high-priority school goals--but growth in these areas is relatively slow, and most contractors will not risk basing reimbursement on the small chance that evidence of growth will be revealed by these criterion tests.

Using judgments of clinically experienced teachers to increase attention to the complexities of performance is considered too subjective (it is not) and too expensive (it is). For all these reasons we can expect some of the complex objectives of instruction to be underemphasized in the typical performance contract testing plan.

The success of Texarkana's first performance-contract year is still being debated. Late-winter (1969-1970) test results looked good, but spring test results were disappointing.* Relatively simple performance items had been used. But the "debate" did not get into that. It started when the project's "outside evaluator" ruled that there had been direct coaching on most, if not all, of the criterion test items. The criterion test items were known by the contractor during the school year. Critics claimed an unethical "teaching for the test." The contractor claimed that both teaching and testing had been directed toward the same specific goals, as should be the case in a good performance contract. The issue is not only one of ethics, it deals with the very definition of education.

*The official evaluation report was written by Andrew and Roberts (1970). Summaries and commentaries have been written by Dyer (1970), Schwartz (1970), and Welch (1970).

Teaching for the Test. Test specialists have recognized an important difference between preparation for a test and direct coaching for a test (Anastasi, 1954, p. 52). To prepare an examinee, the teacher teaches the designated knowledge-skill domain and has the examinee practice good test-taking behavior (e.g., don't spend too much time on overly difficult items; guess when you have an inkling though not full knowledge; organize your answer before writing an essay item) so that relevant knowledge-skill is not obscured. Direct coaching is to teach the examinees how to make correct responses to the specific items on the criterion test.

This is an important difference when criterion test items represent only a small sample of the universe of items representing what has been taught or when the criterion test items are indirect indicators, i.e., correlates, rather than direct measurements, i.e., assessments (see Nunnally, 1959, p. 151).* It ceases to be an important difference when the criterion test is set up to measure directly and thoroughly that which has been taught. In this case, teaching for the test is exactly what is wanted.

The solution of the problem of teaching for the test probably lies in identifying for each objective a very large number (or all) of the items that indicate mastery or progress. Items from standardized tests, if used,** would be included as separate items, not as tests-as-a-whole. The item pool would need to be exhaustive in that, if a student could get a perfect score, there would be no important aspect of the objective that the student would not do well on. A

*The breech also represents the distance between an established teaching profession and challenging instructional technologists.

**Publisher's permission is needed.

separate random sample of items would be drawn for pretest and posttest for each child. Although attractive to a public concerned about the individual child, instructional success would be based on the mean gain of all students of a kind rather than on the gain of individual students. (The use of individual gain scores will be discussed in the next section.) Finding a sufficiently large pool of relevant, psychometrically sound test items is a major chore; but if it can be done, this procedure will prevent "teaching for the test" without introducing a criterion unacceptable to the contractor.

Joselyn (1971) pointed out that the performance contractor and the school should agree in advance as to the criterion procedure though not necessarily to the specific items. To be fair to the contractor, the testing needs to be reasonably close to the teaching. To be fair to the school patrons, the testing needs to be representative of the domain of skills or abilities they are concerned about. A contract to develop reading skills would not be satisfied adequately by gains on a vocabulary test, according to the expectations of most people. All parties need to know how similar the testing is going to be to the actual teaching.

A Dissimilarity Scale. Unfortunately, neither the test specialist nor anyone else has developed scales or grounds for describing the similarity between teaching and testing.* This is a most grievous failing. There is no good way to indicate how closely the tests match the instruction. Complete identity and uniqueness are recognizable by everyone, but important shades of difference are not even presently susceptible to good guessing.

*Richard C. Anderson and his colleagues at the Training Research Laboratory, University of Illinois, have been working on the problem (Anderson, Goldberg, and Hidde, 1971; Wittrock and Hill, 1968).

Some idea of the importance of dissimilarity can be learned from the research literature on transfer of training.

> Working with nonsense syllables, Yum (1931) found that recall memory scores dropped substantially as the test-item stimulus symbol became different from the one learned. He taught persons to say "jury" when he presented the stimulus "toq-bex" and 13 other such stimulus-response combinations. One-third of the learners were retested a day later with the same stimuli; another third were retested with stimuli with one vowel changed; another third with both vowels changed. The results averaged for each subgroup were
>
> 1. Same stimuli on retest: 50% "correct" recall
> 2. Single-letter change: 33% "correct" recall
> 3. Double-letter change: 11% "correct" recall

Generally speaking, as was expected, the greater the dissimilarity, the more difficult the question. In this work and elsewhere (Watts, 1970) another point has been made clear: Small variations can make large changes in item difficulty.

The problem is complicated by the fact that there are many ways for criterion questions to be made dissimilar. Here are some:

1. Syntactic transformation
2. Semantic transformation
3. Change in context or medium
4. Application, considering the particular instance
5. Inference, generalizing from learned instances
6. Implication, adding fast-taught information to generally known information

For examples of some of these transformations, see Table III. Hively, Patterson, and Page (1968) and Bormuth (1970) have discussed procedures for using some of these transformations to generate test items.

Table III

An Example of Transformations
of Information Taught into Test Questions

Information Taught:	Pt. Barrow is the northernmost town in Alaska.
Minimum Transformation Question:	What is the northernmost town in Alaska?
Semantic-Syntactic Transformation Question:	What distinction does Pt. Barrow have among Alaskan villages?
Context-Medium Transformation Question:	The dots on the adjacent map represent Alaskan cities and towns. One represents Pt. Barrow. Which one?
Implication Question:	What would be unusual about summer sunsets in Pt. Barrow, Alaska?

The difficulty of these items depends on previous and intervening learnings as well as the thoroughness of teaching. A considerable difference in difficulty and perceived relevance might be found between the least and most dissimilar questions.* It is apparent that performance contracting in the absence of good information about the similarity between test items and instructional objectives is scarcely an exercise in rationalism.

*The reading items of any contemporary standardized achievement test--as illustrated in Table II--are likely to be more dissimilar to reading teaching (performance contract or regular classroom) than any of the "dissimilarities" shown in Table III.

Analysis of Gain Scores

The following hazards are present in any instruction, not just in performance contracting. The testing specialist sees not one but at least four hazards attendant to the analysis and interpretation of learning scores. They involve (1) grade-equivalent scores, (2) the "learning calendar," (3) the unreliability of gain scores, and (4) regression effects. All show how measures of achievement gain may be spurious. Ignoring any one of them is an invitation to gross misjudgment of the worth of the instruction.

Grade-Equivalent Scores. Standardized achievement tests have the very appealing feature of yielding grade-equivalent scores. Teachers and parents like to use grade-equivalent scores. Raw scores, usually the number of items right, are transformed to scores indicating (for some national reference-group population of students) the grade placement of all students who got this raw score. These transformed scores are called "grade equivalents." The raw scores are not very meaningful to people unacquainted with the particular test; the grade equivalents are widely accepted by teachers and parents. It is probably true that more of them should question the appropriateness of the distribution of scores made by the little-defined reference group as a yardstick for local assessment, but the grade equivalent does represent a piece of test information the public can readily put to use.* Grade equivalents are common terminology in performance contracts.

Unfortunately, grade equivalents are only available from most publishers for tests, not for test items. Thus the whole test needs to be used, in the way

*A shortage of understandable indicators is one reason the schools have not been accountable to the public. However, House (1971) claimed that it is unlikely that educators will use better report procedures even if available because there is much more risk than reward in doing so.

prescribed in its manual, if the grade equivalents are to be meaningful, mean what they are supposed to mean. One problem of using whole tests was discussed in previous sections. Another problem is that the average annual "growth" for most standardized tests is a matter of only a few raw-score points. Consider in Table IV the difference between a grade equivalent of 5.0 and 6.0 with four of the most popular test batteries.

Table IV

Gain in Items Right Needed to Advance
One Grade Equivalent on Three Typical Achievement Tests

	Grade Equivalent		Items Needed To Improve One Year G.E.
	5.0	6.0	
Comprehensive Test of Basic Skills, Level 3: Reading Comprehension	20	23	3
Metropolitan Achievement Test, Intermediate Form B: Spelling	24	31	7
Iowa Tests of Basic Skills, Test A1: Arithmetic Concepts	10	14	4
Stanford Achievement Test, Form W, Intermediate II: Word Meaning	18	26	8

Most teachers do not like to have their year's work summarized by so little a change in performance. Schools writing performance contracts perhaps should be reluctant to sign contracts for which the distinction between success and failure is so small. But to do so requires the abandonment of grade equivalents, at least until a large pool of appropriate items can be identified as to their grade equivalence.*

*Then we would ask, "At what grade level do half the students get this item right?" The score for a student would be the grade equivalents of the most difficult items he passes, with perhaps a correction for guessing.

Instructional specialists (Glaser, 1963; Hively, Patterson, and Page, 1968) have questioned the appropriateness of grade equivalents or any other "norm referencing" for interpreting items. They object to defining performance primarily by indicating who else performs this well. Clearly the items on all standardized tests have been selected on the basis of their ability to discriminate between the more and less sophisticated students rather than as to whether or not they tell whether or not a person has mastered his task. Joselyn (1971) said that the items left may do the poorest job of describing performance. Jackson (1970) summarized the research and writing of those who endorse only those standardized test items which directly indicate successful attainment of the instructional objectives. But the items Jackson's authors would like educators to use usually do not exist--or if they do, there whereabouts are unknown. Creating and field-testing new test items is a difficult, time-consuming, costly task. For a local performance contract, the cost of developing "their own" criterion items could easily exceed the entire cost of instruction. In the years ahead, such criterion items must become available for purchase. Grade equivalents, as Lennon (1971) concluded, in spite of their apparent utility, are too gross for the measurement of individual short-term learning.

The School Calendar. For most special instructional programs in the schools, criterion tests will be administered at the beginning of and immediately following instruction, often in the first and last weeks of school. There is a large amount of distraction in the schools those weeks, but choosing other times for pre- and post-testing has its hazards too. Getting progress every several weeks during the year is psychometrically preferred (Wick and Beggs, 1971); but most instructional people are opposed to "all that testing."

Children learn year round, but the evidence of learning that gets inked on pupil-personnel records comes in irregular increments from season to season. Winter is the time of most rapid academic advancement, summer the least. Summer, in fact, is a period of setback for many youngsters. Beggs and Hieronymus (1968) found punctuation skills to spurt more than a year's worth between October and April but to drop almost half a year between May and September. Discussing their reading test, *Gates and MacGinitie* (1965) said,

> "...in most cases, scores will be higher at the end of one grade than at the beginning of the next. That is, there is typically some loss of reading skill during the summer, especially in the lower grades."

The picture will be different, of course, depending on what the learners do in and out of school.*

The first month or two of the fall, when students first return to school, is the time for getting things organized and restoring general skill abilities lost during the summer. According to some records, spring instruction competes with only partial success with other spring attractions. Thus, the learning year is a lopsided year, a basis sometimes for miscalculations. Consider the results of testing shown in Table V.

Table V

Learning Calendar for a Typical Fifth-Grade Class

	Month								
	S	O	N	D	J	F	M	A	M
Mean Achievement Score	5.0		5.3		5.6	5.9		6.2	6.3

*A spring slowdown and summer setback sometimes occur in conventional school programs. If the instructional program began in March or in June, the results would not necessarily be the same.

The every-two-months averages in Table V are fictitious, but they represent test performance in a typical classroom. The growth for the year appears to be 1.3. No acknowledgement is made there that early-September standardized test results were poorer than those for the previous spring. For this example the previous May mean (not shown) was 5.2. The real gain, then, for the year is 1.1 grade equivalents rather than the apparent 1.3. It would be inappropriate to pay the contractor for a mean gain of 1.3.

Another possible overpayment on the contract can result by holding final testing early and extrapolating the previous per-week growth to the weeks or months that follow. In Texarkana, as in most schools, spring progress was not as good as winter. If an accurate evaluation of contract instructional services is to be made, repeated testing, perhaps a month-by-month record* of learning performances needs to be considered.

Perhaps the biggest when-to-test problem arises from the common belief that schooling is not supposed to aim at terminal performance (at project's end) but to aim at continuing performance in the weeks and months and years that follow. Many diverse instructional specialists (Gagné, Mayor, Garstens, and Paradise, 1962; Traub, 1966; Atkin, 1963) agree that the instructor should use different tactics to maximize long-term rather than short-term gain. Teachers are inclined to emphasize long-term aims; the performance contractor has proposed to deal with short-term aims. They will disagree about the allocation of teaching time. The contractor points out that he is there because the school recognized

*Wrightman and Gorth (1969) described Project CAM as a model for a continuous (perhaps every two weeks) performance monitoring record.

that some students need immediate remedial work. He, the contractor, is not going to dilute his remedy just because there are many other important objectives for the school. He is not going to give major attention to how this instruction will coordinate with what the student will get in simultaneous and subsequent instruction. His is a defensible position. Whether or not he should be placed in a position that will substantially reduce emphasis on long-range educational goals is an issue needing attention early in any discussions about performance contracting.

Unreliable Gain Scores. Most performance contracts pay off on an individual-student basis. The contractor may be paid for each student who gains more than an otherwise expected amount. This practice is commendable in that it emphasizes the importance of each individual learner and makes the contract easier to understand, but it bases payment on a precarious landmark: the gain score.

Let us see how unreliable the performance-test gain score is. For a typical standardized achievement test with two parallel forms, A and B, we might find the following characteristics reported in the test's technical manual:

Reliability of Test A = +.84
Reliability of Test B = +.84
Correlation of Test A with Test B = +.81

Almost all standardized tests have reliability coefficients at this level. By using the standard formula (Thorndike and Hagen, 1969, p. 197), we find a disappointing level of reliability for the measurement of improvement.

Reliability of Gain Scores (A - B or B - A) = +.16

The manual would indicate the raw score and grade-equivalent standard deviations. For one widely used test they are 9.5 items and 2.7 years, respectively. Using these values we can calculate the errors to be expected. On the average, a student's...

...raw score would be in error by 2.5 times
...grade equivalent would be in error by 0.72 years
...grade-equivalent gain score would be in error by 1.01 years

The error is indeed large.

Consider what this means for the not-unusual contract whereby the student is graduated from the program, and the contractor is paid for his instruction, on any occasion that his performance score rises above a set value. Suppose--with the figures above--the student exits whenever his improvement is one year grade equivalent or better. Suppose also, just to make this situation simpler, that there is no intervening training and that the student is not influenced by previous testing. Here are three ways of looking at the same situation:

> Suppose that a contract student were to take a different parallel form of the criterion test on three successive days immediately following the pretest. The chances are better than 50:50 that on one of these tests the student would have gained a year or more in performance and would appear to be ready to graduate from the program.
>
> Suppose that three students were to be tested with a parallel form immediately after the pretest. The chances are better than 50:50 that one of the three students--entirely due to the errors of measurement--would have gained a year or more and appear ready to graduate from the program.
>
> Suppose that 100 students were admitted to contract instruction and pretested. After a period of time involving no training, they were tested again and the students gaining a year were graduated. After another period of time, another test and another graduation. After the fourth terminal testing, even though no instruction had occurred, the chances are better than 50:50 that two-thirds of the students would have been graduated.

In other words, the unreliability of gain scores can give the appearance of learning that actually does not occur.

The unreliability also will give an equal number of false impressions of deteriorating performance. These errors (false gains and false losses) will

balance out for a large group of students. If penalties for losses are paid by the contractor at the same rate bonuses are paid for gains, the contractor will not be overpaid. But according to the way contracts are being written, typified in the examples above, the error in gain scores does not balance out; it works in favor of the contractor. Measurement errors could be capitalized upon by unscrupulous promoters. Appropriate checks against these errors are built into the better contracts.

Errors in individual gain scores can be reduced by using longer tests. A better way to indicate true gain is to calculate the discrepancy between actual and expected final performances.* Expectations can be based on the group as a whole or on an outside control group. Another way is to write the contract on the basis of mean scores for the group of students.** Corrections for the unreliability of gain scores are possible, but they are not likely to be considered if the educators and contractors are statistically naive.

Regression Effects. Buried back here in this paper is probably the source of the greatest misinterpretation of the effects of remedial instruction. Regression effects are easily overlooked but need not be; they also are susceptible to correction. For any pretest score the expected regression effect can be calculated. Regression effects make the poorest scorers look better the next time tested. Whether measurements are error-laden or error-free, meaningful or

*Tucker, Damarin, and Messick (1965) have discussed change scores that are independent of and dependent on the initial standing of the learner. A learning curve fitted to test scores could be used to counter the unreliability of individual scores.

**This would have the increased advantage of discouraging the contractor from giving preferential treatment within the project to students who are in a position to make high pay-off gains.

meaningless, when there is differential change between one measurement occasion and another (i.e., when there is less-than-perfect correlation), the lowest original scorers will make the greatest gains and the highest original scorers will make the least. On the average, posttest scores will, relative to their corresponding pretest scores, lie in the direction of the mean. This is the regression effect. Lord (1963) discussed this universal phenomenon and various ways to set up a proper correction for it.

The demand for performance contracts has occurred where conventional instructional programs fail to develop--for a sizable number of students--minimum competence in basic skills. Given a distribution of skill test scores, the lowest-scoring students, ones most needing assistance, are identified. It is reasonable to suppose that under unchanged instructional programs they would drop even further behind the high-scoring students. If a retest is given, however, after any period (of conventional instruction, of special instruction, or of no instruction), these students will no longer be the poorest performers. Some of them will be replaced by others who then appear to be most in need of special instruction. Instruction is not the obvious influence here--regression is. Regression effect is not due to test unreliability--but it causes some of the same misinterpretations. The contract should read that instruction will be reimbursed when gain exceeds that attributable to regression effects. The preferred evaluation design would call for control group(s)* of similar students to provide a good estimate of the progress the contract students would have made in the absence of the special instruction.

*Wardrop (1971) has discussed the problem of control groups that do not provide an appropriate control.

The Social Process

The hazards of specific performance testing and performance contracting are more than curricular and psychometric. Social and humanistic challenges should be raised too. The teacher has a special opportunity and obligation to observe the influence of testing on social behavior.

At several places in the preceding pages I have referred to the uniqueness of making major personal and scholastic decisions on the sole basis of student performances. This is unique also because it puts the student in a position of administrative influence. Here he can influence what the instructional benefit would look like. He can make it look better or poorer than it really is (Anastasi, 1954, p. 56). More responsibility for school control possibly should accrue to students, but performance contracts seem a devious way to give it.

Even if he is quite young, the student is going to be aware that his good work will bring rewards to the contractor. Sooner or later he is going to know that, if he tests poorly at the beginning, he is able to do more for himself and the contractor. Bad performances are in his repertoire--he may be more anxious to make the contractor look bad than to make himself look good. He may be under undue pressure to do well on the posttests. These are pupil-teacher interactions that should be watched carefully.

To motivate the student to learn and to make him want more contract instruction, many contractors use material or opportunity-to-play rewards. (Dorsett used such merchandise as transistor radios.) Other behavior modification strategies (Meecham and Wiesen, 1969) are common. The proponents of such strategies argue that, once behavior has been oriented to appropriate tasks, the students can gradually be shifted from extrinsic rewards to intrinsic. That they

can be shifted is probably true, that that it will happen without careful deliberate work by the instructional staff is unlikely. It is not difficult to imagine a performance-contract situation in which the students become even less responsive to the rewards of conventional instruction than they were before.

Still another hazard of performance contracting and many other uses of objectives and test items is that by using them as we do, without acknowledging how much they indirectly and incompletely represent educational goals, we misrepresent education. People inside school and out pay attention to grades and tests and monetary reimbursements. We may not value factual knowledges and simple skills proportionately to the attention they get, but we have ineffective ways of indicating what our priorities really are (Stake, 1970).

It is difficult for many people to accept the fact that in conventional classrooms a vast number of educational goals are simultaneously pursued (Gooler, 1971). Efforts to get teachers to specify those objectives result in a simplified and incomplete list. The performance contractor has an even shorter list. Even if performance contracting succeeds in doing the relatively small job it aims to do, adequate arguments have not been made that this job should be given the priority and resources that the contractors require.

In early 1971 performance contracting appears to be popular in Washington with the current administration because it encourages the private-business sector to participate in a traditionally public responsibility. It is popular among some school administrators because it gets some tough-to-get federal funds, because it is a novel and possibly cheaper way to get new talent working on old problems, and because the administrator can easily blame the outside agency and the government if the contract instruction is unsuccessful. It is unpopular with the American Federation of Teachers because it reduces the control the Union

has over school operations and it reduces the teacher's role as a chooser of what learnings students are most in need of. It is popular among most instructional technologists because it is based on a number of well-researched principles of teaching and because it enhances their role in school operations.

The accountability movement as a whole is likely to be a success or failure on such socio-political items as in the foregoing oversimplified list. Cohen (1970) reminded evaluators to look for the issues the decision makers are concerned about. All too seldom do these include the measures of performance considered in this paper. The measurement of the performance of "performance contracting" is an even more hazardous procedure than the measurement of student performances.

Summary. Without yielding to the temptation to harass new efforts to provide instruction, educators should continue to be apprehensive about evaluating teaching on the basis of performance testing alone. They should know how difficult it is to represent educational goals with statements of objectives. They should know how costly it is to provide suitable criterion testing. They should know that the common-sense interpretation of these results is frequently wrong but that many members of the public and the profession think that special designs and controls are extravagant and mystical.

Performance contracting emerged because people inside the schools and out were dissatisfied with the instruction some children are getting. Implicit in the contracts is the expectation that available tests can measure the newly promised learning. The standardized test alone cannot measure the specific outcomes of an individual student with sufficient precision. This limitation and other hazards of performance measurement are applicable, of course, to the measurement of specific achievement in regular school programs.

Bibliography

(Not Cited)

The following items are relevant to this topic but were not cited in this paper:

Blaschke, Charles. Performance contracting in education. Educational Turnkey Systems. Washington, D. C., March 1970.

Dyer, Henry S., Linn, Robert L., and Patton, Michael J. A comparison of four methods of obtaining discrepancy measures based on observed and predicted school system means on achievement tests. American Educational Research Journal, 1969, 6, 591-605.

Elam, Stanley. The age of accountability dawns in Texarkana. Phi Delta Kappan, 1970, 15(10), 509-514.

Linn, Robert L., Rock, Donald A., and Cleary, T. Anne. The development and evaluation of several programmed testing methods. Educational and Psychological Measurement, 1969, 29, 129-146.

Phi Delta Kappa Commisson on Evaluation. Educational evaluation and decision making. Itasca, Illinois: Peacock Publishers, 1971.

Popham, W. James and Husek, Theodore R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.

Scriven, Michael. The methodology of evaluation. In Robert E. Stake (Ed.), Perspectives of curriculum evaluation, AERA Monograph Series on Curriculum Evaluation, No. 1. Chicago: Rand McNally, 1967. Pp. 39-83.

Stake, Robert E. The countenance of educational evaluation. Teachers College Record, 1967, 68, 523-540.

Tyler, Ralph W. Educational evaluation: new roles, new means, Sixty-eighth Yearbook, National Society for the Study of Education. Chicago: University of Chicago Press, 1969.

Webb, Harold V. Performance contracting: is it the new tool for the new boardmanship? American School Board Journal, 1970, 185(5), 28-36.

(Cited)

Analysis of Learning Potential, Advanced I Battery, Form A. George A. Prescott, Coordinating Editor. New York: Harcourt Brace Jovanovich, 1970.

Anastasi, Anne. Psychological testing. New York: Macmillan, 1954.

Anderson, Richard C., Goldberg, Sheila M., and Hidde, Janet L. Meaningful processing of sentences. Journal of Educational Psychology, 1971, in press.

Andrew, Dean C. and Roberts, Lawrence H. Final evaluation report on the Texarkana Dropout Prevention Program, Magnolia, Arkansas: Region VIII. Education Service Center, July 20, 1970. (mimeo)

Atkin, J. Myron. Behavioral objectives in curriculum design: a cautionary note. The Science Teacher, May 1968, 27-30.

Atkin, J. Myron. Research styles in science education. Paper presented at the annual meeting of the National Association for Research in Science Teaching, Chicago, February 1968.

Atkin, J. Myron. Some evaluation problems in a course content improvement project. Journal of Research in Science Teaching, 1963, 1, 129-32.

Baker, Eva L. Experimental assessment of the effects of the Probe System. Paper presented at the annual meeting of the American Educational Research Association, Minneapolis, March 1970.

Beggs, Donald L. and Hieronymus, Al. Uniformity of growth in the basic skills throughout the school year and during the summer. Journal of Educational Measurement, 1968, 5, 91-97.

Bhaerman, Robert. A paradigm for accountability. American Teacher, 1970, 55(3), 18-19.

Bloom, Benjamin S., Englehart, Max D., Furst, Edwin J., Hill, William H., and Krathwohl, David R. A taxonomy of educational objectives: Handbook I, the cognitive domain. New York: David McKay, 1956.

Bormuth, John. On the theory of achievement test items. Chicago: University of Chicago Press, 1970.

Cohen, David K. Politics and research--evaluation of large-scale education programs. Review of Educational Research, 1970, 40(2), 213-238.

Cronbach, Lee J. Validation of educational measures. Proceedings of the 1969 Invitational Conference on Testing Problems. Princeton, New Jersey: Educational Testing Service, 1969. Pp. 36-52.

Dyer, Henry S. Performance contracting: too simple a solution for difficult problems. The United Teacher, November 29, 1970, 19-22.

Ebel, Robert L. When information becomes knowledge. Science, 1971, 171, 130-131.

Gagné, Robert M. Curriculum research and the promotion of learning. In Robert E. Stake (Ed.), Perspectives of curriculum evaluation, AERA Monograph Series on Curriculum Evaluation, No. 1. Chicago: Rand McNally, 1967. Pp. 19-38.

Gagné, R. M., Mayor, J. R., Garstens, Helen L., and Paradise, N. E. Factors in acquiring knowledge of a mathematical task. Psychological Monographs, 1962, 76(7, Whole No. 526).

Gates, Arthur I. and MacGinitie, Walter H. Technical manual for the Gates-MacGinitie Reading Tests. New York: Teachers College Press, Columbia University, 1965. P. 5.

Glaser, Robert. Instructional technology and the measurement of learning outcomes: some questions. American Psychologist, 1963, 18, 519-521.

Gooler, Dennis D. Strategies for obtaining clarification of priorities in education. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign, 1971.

Goslin, David A. Ethical and legal aspects of the collection and use of educational information. Paper presented at the Invitational Conference on Testing Problems, New York, October 1970.

Grobman, Hulda. Developmental curriculum projects: decision points and processes. Itasca, Illinois: Peacock Publishers, 1970.

Hastings, J. Thomas, Runkel, Philip J., and Damrin, Dora E. Effects on Use of Tests by Teachers Trained in a Summer Institute, Cooperative Research Project No. 702. Urbana: Bureau of Educational Research, College of Education, University of Illinois, 1961.

Hively, Wells II, Patterson, Harry L., and Page, Sara H. A "universe-defined" system of arithmetic achievement tests. Journal of Educational Measurement, 1968, 5(4), 275-290.

Hoffman, Banesh. The tyranny of testing. New York: Collier Books, 1962.

Holt, John. The underachieving school. New York: Pitman, 1969.

House, Ernest R. The conscience of educational evaluation. Paper presented at the Ninth Annual Conference of the California Association for the Gifted, Monterey, February 1971.

Jackson, Rex. Developing criterion-referenced tests. Princeton, New Jersey: ERIC Clearing House on Tests, Measurement, and Evaluation, Educational Testing Service, June 1970.

Jenkins, Joseph R. and Deno, Stanley L. Effects of instructional objectives on learning. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, February 1969.

Joselyn, E. Gary. Performance contracting: what it's all about. Paper presented at Truth and Soul In-Teaching Conference of the American Federation of Teachers, Chicago, January 1971.

Krathwohl, David R. AERA Presidential Address. Presented at the annual meeting of the American Educational Research Association, Los Angeles, February 1969.

Lennon, Roger T. Testimony of Dr. Roger T. Lennon as expert witness on psychological testing. New York: Harcourt, Brace, and World, no date.

Lennon, Roger T. Accountability and performance contracting. Paper presented at the annual meeting of the American Educational Research Association, New York, February 1971.

Lessinger, Leon. Engineering accountability for results in public education. Phi Delta Kappan, 1970, 52(4), 217-225.

Lieberman, Myron. An overview of accountability. Phi Delta Kappan, 1970, 52(4), 194-195.

Lindquist, E. F. (Ed.) Educational measurement. Washington: American Council on Education, 1951.

Lord, Frederic M. Elementary models for measuring change. In Chester W. Harris (Ed.), Problems in measuring change. Madison: University of Wisconsin Press, 1963. Pp. 21-38.

Lord, Frederic. A theory of test scores. Psychometric Monograph Number 7, The Psychometric Society. Philadelphia: George E. Ferguson Co., 1952.

Lortie, Dan C. National decision-making: is it possible today? The EPIE Forum, 1967, 1, 6-9.

Mager, Robert F. Preparing objectives for programmed instruction. San Francisco: Fearon Press, 1962.

McGhan, Barry R. Accountability as a negative reinforcer. American Teacher, 1970, 55(3), 13.

McNeil, John D. Concomitants of using behavioral objectives in the assessment of teacher effectiveness. Journal of Experimental Education, 1967, 36.

Meacham, Merle L. and Wiesen, Allen E. Changing classroom behavior: a manual for precision teaching. Scranton: International Textbook Co., 1969.

National School Board Journal Staff. Two out of three boardmen buy performance contracting. National School Board Journal, November 1970, 35-36.

Nunnally, Jum C. Jr. Tests and measurements: assessment and prediction. New York: McGraw-Hill, 1959.

Oettinger, Anthony G. Run, computer, run. Cambridge, Massachusetts: Harvard University Press, 1969.

Popham, W. James. Objectives and instruction. In Robert E. Stake (Ed.) Instructional objectives, AERA Monograph Series on Curriculum Evaluation, No. 3. Chicago: Rand McNally, 1969. Pp. 32-52.

Schrag, Peter. End of the impossible dream. Saturday Review, 1970, 53(38), 68-70, 92-96.

Schwartz, Ronald. Performance contracting: industry's reaction. The Nation's Schools, 1970, 86, 53-55.

Silberman, Charles E. Crisis in the classroom: the remaking of American Education. New York: Random House, 1970.

Sizer, Theodore R. Social change and the uses of educational testing: an historical view. Paper presented at the Invitational Conference on Testing Problems, New York, October 1970.

Stake, Robert E. Objectives, priorities, and other judgment data. Review of Educational Research, 1970, 40, 181-212.

Thorndike, E. L. The measurement of educational products, Seventeenth Yearbook of the National Society for the Study of Education, Part II. Bloomington, Illinois: Public School Publishing Co., 1918. Also in Geraldine M. Joncich (Ed.), Psychology and the science of education. New York: Bureau of Publications, Teachers College, Columbia University, 1962. P. 151.

Thorndike, R. L. Educational psychology, volume 1: the original nature of man. New York: Teachers College, Columbia University, 1921. Pp. 11-12.

Thorndike, R. L. and Hagen, Elizabeth. Measurement and evaluation in psychology and education. (3rd ed.) New York: Wiley, 1969.

Traub, Ross E. Importance of problem heterogeneity to programmed instruction. Journal of Educational Psychology, 1966, 57, 54-60.

Tucker, Ledyard R, Damarin, Fred, and Messick, Samuel. A base-free measure of change, Research Bulletin RB-65-16. Princeton, New Jersey: Educational Testing Service, 1965.

Tyler, Ralph W. Basic principles of curriculum and instruction. Chicago: University of Chicago Press, 1950. P. 83.

Wardrop, James L. Some particularly vexing problems in experimentation on reading. Reading Research Quarterly, Spring 1971, in press.

Watts, Graeme H. Learning from prose material: effects of verbatim and "application" questions on retention. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign, 1970.

Wrightman, L. and Gorth, W. P. CAM: the new look in classroom testing. Trend, Spring 1969, 56-57.

Welsh, James. D. C. perspectives on performance contracting. Educational Researcher, Volume XXXI, October 1970, 1-3.

Wick, John and Beggs, Donald L. Evaluation for decision-making in the schools. New York: Houghton Mifflin, 1971.

Wittrock, M. C. and Hill, Claude E. Children's preferences in the transfer of learning. Final Report, Project No. 3264, U. S. Department of Health, Education, and Welfare, November 1968.

Yum, K. W. An experimental test of the law of assimilation. Journal of Experimental Psychology, 1931, 14, 68-82.

Zahorik, John A. The effect of planning on teaching. The Elementary School Journal, December 1970, 143-151.